Technology Watch Report

The large-scale archival storage of digital objects

Jim Linden, Sean Martin, Richard Masters, and Roderic Parker
The British Library

DPC Technology Watch Series Report 04-03
February 2005

Contents

Summary

This report is intended to be of use to any digital repository that is looking at long-term archival storage issues, especially on a large scale.

The British Library has national responsibilities for preserving and providing access to digital materials for the very long term. Many of these materials will be acquired under Legal Deposit legislation. However, the Library's requirements are not unlike those of many large digital archives, and the case study presented here has wider application.

This report describes the issues surrounding storage technology from the perspective of a digital repository. It covers the British Library's thinking that led to conclusions about the design of a system for the preservation storage of digital objects, and the specific conclusions about the preservation storage architecture to adopt. The planning figure for the storage required was that it would grow to over 100 Terabytes (100 thousand gigabytes, or 100 million megabytes) in 3 years.

A multi-site architecture was selected. Each site would provide, amongst other functions, a software layer providing storage services that was independent of the actual physical storage. This physical storage would use commodity storage units from a number of vendors, which could be replaced over time without having to alter any software. The storage initially purchased uses the relatively new Serial ATA standard.

In the course of this work over 30 UK storage suppliers were interviewed, which gave a good view of the products and services available in the UK at the time, as well as of likely trends in the market and their future impact on the system.

The work reported on was conducted between November 2003 and October 2004.

General issues for digital object storage

Need for a large and secure data storage system

Storing digital objects leads, if nothing else, to large byte-counts. Long gone are the days of 'a file on a floppy'. Individuals archiving their personal data are already working with gigabytes of data, backed up onto external optical or magnetic memory devices (such as 4.7 Gigabyte (GB) DVDs or 4 GB flash memory devices). An institution's digital repository (which may have a label such as 'library' or 'archive') may include terabytes of data spread across thousands of objects in hundreds of formats. It may, for ease of handling and control, want to transfer the contents to a more readily managed medium. The sheer ease of producing and proliferating electronic files, combined with the perception that storage is so cheap as not to be a concern at all, leads to an uncontrolled explosion in numbers and storage volumes.

Security of the collections has always been a concern for a library. No responsible library fails to record where its stock should be, or to preserve its stock for what is seen as its useful lifetime. Electronic objects bring in new problems and issues, such as

- where is the object?
- who has access to any one object?
- who can copy or change it?
- are there several copies of the object and, if so,
    - do they serve different purposes?
    - are any differences from the original object significant?
- what is the lifetime history of the changes made?
- is there a reliable way to prevent copies or changes from being made?

These twin drivers of size and security lead to the conclusion that the digital preservation of any but a trivial number of objects for anything but a short period of time will of necessity involve the purchase and management of a large and secure data storage system.

Total Cost of Ownership

The 'Total Cost of Ownership' (TCO) of storage (which includes the life-long cost of its management) varies from organisation to organisation, but can only be minimised by starting from understanding the hidden costs of unmanaged storage as well as the exposed costs. Gartner research [1] indicates that for some storage technologies, capital costs amount to only ⅓ of the total lifetime costs. Other unattributed research [2] puts capital cost at little more than 10% of the overall cost.

Those who do not have a good understanding of storage TCO tend to focus on hardware acquisition costs, but a typical lifecycle, as shown in Figure 1 (which is illustrative and makes no attempt to be quantitative) will involve other costs. These can include

- Planning
    - process planning
    - deciding evaluation criteria
    - product evaluation
    - selection
- Acquisition
    - direct purchase costs
    - installation planning
    - pre-installation training
    - site preparations
- Operation and maintenance
    - staffing
    - staff training

---

[1] Quoted in *The Hidden Costs of Unmanaged Storage,* Leslie Wood, July 7, 2003. www.enterprisestorageforum.com/management/features/article.php/2231701

[2] *Virtual Storage -- is it there yet?*, Rupert Goodwins, ZDNet UK, May 29, 2002. http://insight.zdnet.co.uk/hardware/servers/0,39020445,2111098,00.htm

- configuration
- testing prior to data loading
- monitoring
- consumables
- backup
- data replication
- integration with third-party software
- multi-vendor integration and pre-testing
- reporting
- downtime
- mid-life refurbishment
- end-of-life planning
- data transfer to successor storage

- Disposal
    - secure data deletion
    - physical removal
    - disposal costs (landfill or charity)

Full consideration of the Total Cost of Ownership of storage is clearly an issue that no organisation can afford to neglect.

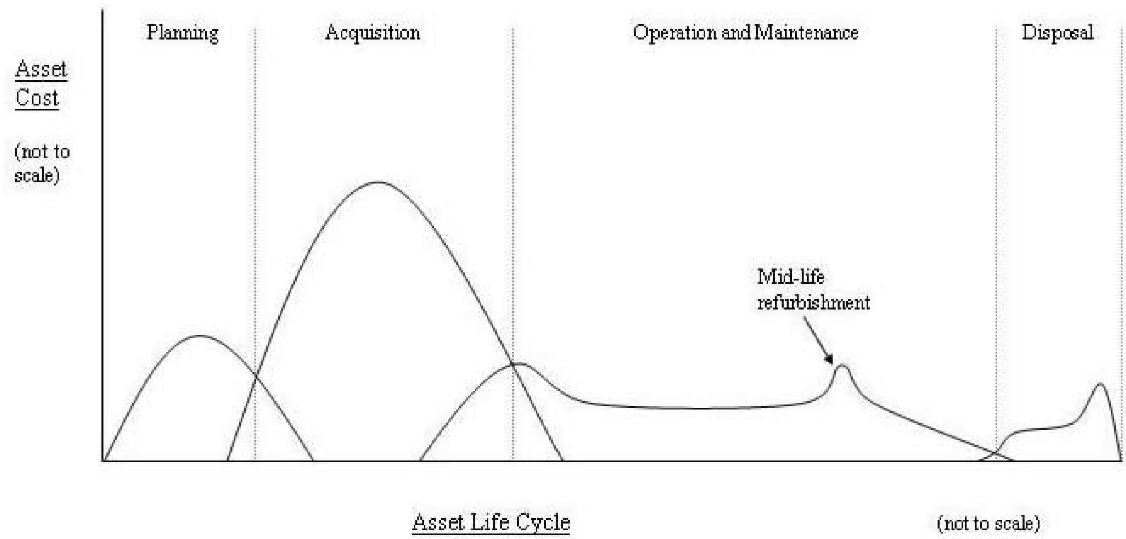

Figure 1. Typical hardware lifecycle

Storage technology issues

Technical storage background

Like any other technology, data storage has never stood still. Observations over five or more years have led to two trends being seen almost as 'laws':

- The capacity of storage media doubles every 12 to 18 months
- The unit purchase cost of storage reduces by 30% to 40% each year

Storage technologies

The two principal technologies for large-scale storage have for some years been magnetic disk drives (primarily for data needed quickly) and tape drives (primarily for archive and backup storage). The perceived distinction between when the two different technologies should be used is becoming blurred. Capacities for both technologies have been increasing, but the difference in cost per gigabyte has been decreasing, making disk technology a more attractive proposition for archival or off-line storage where once it would never have been considered.

A storage technology that has matured significantly in the past year is Serial ATA (Serial Advanced Technology Attachment). [3] Version 1.0 of the specification was published in August 2001, by which time work leading to the 2004 version 2.0 had already started.

Serial ATA is an interface standard: it defines how a storage device is linked (indirectly) to the processor unit of a computer [4] and so also defines the speed and efficiency of data transfer to and from the storage system. Serial ATA grew out of Parallel ATA (also called IDE, an abbreviation for either Intelligent Drive Electronics or Integrated Drive Electronics). Although it had a background in personal computers, storage vendors realised that it offered many advantages for larger-scale storage systems. [5]

Another recent development in the field of faster and more reliable storage interfaces has been Serial Attached SCSI drives. These promise to offer better availability and performance, and to be better suited to mainstream enterprise applications.[6] However, at the time of the work reported here (June 2004) we were not aware of any commercial products.

Serial ATA offers lower costs per gigabyte and higher capacities per drive than Serial Attached SCSI, and is seen as better suited to longer-term file storage.[7] Part of the cost difference comes from the greater degree of control needed for a SCSI disk spinning at between 10,000 and 15,000 rpm, compared with an ATA disk spinning typically at 5,400 or 7,200 rpm.

To give an idea of the changes in costs, in 1996 the raw purchase cost per gigabyte of a (parallel) ATA disk drive was over $100, while the cost per gigabyte of tape drive media was under $10. Seven or so years later, ATA disk drives cost less than $2 per gigabyte, while the media for tape drives costs about $1 per gigabyte.[8]

Magnetic tapes and disks have been being developed since the 1950s, but optical drives have been in use only since the 1980s. Their performance is usually seen as being worse than that of magnetic media, though their lifespan is longer. While optical disks are a dominant storage medium in the consumer electronics market, they remain a niche component in the business storage market (not helped by the number of different standards that emerged initially), with about 4% of sales by value worldwide. Well over 90% of the world's digital data is thought to be stored magnetically.

Virtualisation

'Virtualisation' is the name given to techniques of combining disparate physical storage devices and presenting them to the operating system as a logical whole. [9] It offers ways of operating increasingly diverse physical storage units, though not without its own extra costs.

Network technologies

Quite apart from the competing technologies trying to squeeze more data on to different media, and to improve the read and write access speeds to that data, there is also the issue of organizing that media within a computer system. Broadly speaking there are three distinct ways of doing this:

---

[3] The 'AT' in 'ATA' is the same 'AT' as in the IBM PC AT personal computer, introduced in August 1984. This was a version of the IBM PC, with an Intel 80286 processor, a 16-bit bus, a medium-speed 20 megabyte hard disk, and a 1.2 megabyte floppy disk drive. Times change.

[4] For further background and description see for example *Serial ATA. A promising new alternative for enterprise storage applications*, Sept 2002. ftp://download.intel.com/technology/serialata/pdf/25179001.pdf

[5] Intuitively it seems odd that a serial transfer method is replacing a parallel transfer one, but serial transfer is far easier to control technically, and has a number of other advantages.

[6] See, for example, *ANSI agrees on Serial Attached SCSI standard*, Clint Boulton, 2004. http://www.enterprisestorageforum.com/technology/news/article.php/3305441

[7] See, for example, *Serial ATA - It's Time to Get in Line*, John Vacca, 2003. http://www.enterprisestorageforum.com/hardware/features/article.php/1482351

[8] Figures quoted from *What drives your data center*, Kevin Gray (Maxtor Corp.), Computer Technology Review, February 2003.

[9] For a brief description see for example *What is storage virtualisation?*, 8th November 2004. http://www.expresscomputeronline.com/20041108/technology05.shtml

- Direct Attached Storage (DAS), where the storage devices are attached to mainframes or individual servers. Access is available only to those also attached to these machines
- Network Attached Storage (NAS), where specialised storage devices are attached to an existing network. Any client attached to the network can access these devices.
- Storage Area Network (SAN), where the storage devices are attached to a network of their own, which is in turn connected to the main network. Again, any client attached to the main network can access the storage, but the distinct nature of the separate storage network allows for specialised management both of the data network and of the main (applications) network.

Storage units used in the NAS architecture tend to be desktop-rated rather than server-rated units, neither designed nor built for round-the-clock operation. This leads to their being less robust. Some vendors have tried to work round this with products which combine the SAN approach with the NAS approach.

Interoperability within storage networks, caused by the lack of mature standards, remains a problem.

Transport technologies

Whilst there are several technologies and standards relating to data transfer within and between networks and systems, only two groups are of major significance:

- Fibre Channel, used for Storage Area Networks. This can be difficult to set up and maintain for the extremely fast speeds that can be achieved.
- Internet Protocol (IP), used in older Ethernet-based networks and newer Gigabit Ethernet networks. This is most often manifested as
  - Fibre Channel over Internet Protocol (FCIP) and
  - Internet Small Computer Systems Interface (iSCSI)

The IT Storage Market [10]

The IT storage market has traditionally been served by two distinct kinds of vendor, serving two distinct markets.

- Service providers. These vendors provide high-quality solutions, with cutting-edge technology that is often highly innovative. What they offer is sometimes called 'enterprise level storage': the storage volumes and the service to support them are pitched at the enterprise or corporate level, with premium prices to match. The customer pays for and receives a high level of service, though often having to be locked in to one vendor. Service features include:
  - maximised performance,
  - maximised resilience with respect to individual hardware failures, but typically within equipment hosted at a single location, and
  - lowered operational costs for enterprise solutions, where, for example, data is dynamic and storage solutions serve the needs of multiple users.
- Storage solution vendors. These vendors also offer high quality solutions, but they use hardware that is not right at the leading edge and follows more established technical standards. Their solutions offer less in the way of storage management, giving the customer more to do in terms of load balancing, upgrading applications, adding extra storage, handling bottlenecks, and dealing with failures. To some extent the storage units are interchangeable, hence the term 'commodity storage' being applied to them.

This distinction is not as clear as it used to be: increased capacities and much improved storage management software are seeing to that. However, it seems to be a distinction that the vendors themselves are happy to keep making.

It has been suggested [11] that storage hardware will all be seen as a commodity item within a few years, since increased hardware reliability allied to wider acceptance of standards (and hence greater interoperability) will lead to added-value software services being the only factor to differentiate between vendors. But it can also be argued that it is the very presence of this extra software that distinguishes storage service providers from commodity storage vendors.

---

[10] A useful market research report is *Mass storage systems: developments and trends*, Patricia Cloar, Business Communications Co.; Norwalk, Conn.; 2003. Business Opportunity Report GB-252. ISBN 1 56965 9532.

[11] See for example *Storage hardware set for commodity status*, M Courtney (Meta Group), June 2004. http://www.vnunet.com/print/1156277

Demand for increased storage capacity

Demand for increased storage capacity comes from a number of sources. These include: [12]

- Increased regulation of business by government, leading to legal requirements to archive business data for compliance purposes (including internal emails) for set numbers of years.
- Increased demands for product development data (such as drug trial data and engineering test data) to be retained for many years (not necessarily for a defined period)
- Increased demands for product maintenance data to be retained (e.g. aircraft maintenance records)
- The development of decision support, customer relationship, point of sale, and other data-hungry systems
- Major growth in data intensive fields of study such as climate modelling and earth imaging.

Data generated by these demands all has a finite lifetime. Even if this lifetime is not clear, it is measurable in decades. However, an academic library dealing with its digital data may well think in terms of centuries (indeed in the British Library we talk of "forever"). 'Normal' practices in business and commerce may not apply.

Storage Management

One side-effect of the increasing volume of data being stored is that its management is becoming more active. There are two general approaches, though terminology isn't always used precisely in the industry:

- Hierarchical Storage Management (HSM) places files in a storage system which contains at least two different kinds of storage. Business rules are set up to govern automatic file placement: in general terms, the newer or the more recently used files are placed on the drives with the fastest access, and older &/or less used files are placed on lower-speed, lower-cost, drives.
- Data or Information Lifecycle Management (ILM) is a newer development. File placing is automatic, but is driven by the policy and business values of the data itself. Content management is an essential component, though in practical terms it has proved difficult so far to integrate it with storage management.[13]

Another business-driven change is that the drive towards 24/7 working has reduced or eliminated the traditional 'silent hours' when storage housekeeping tasks (backups etc.) could be carried out. The demand now is for round-the-clock working, 'always-on' networks, and the replacement of faulty or old storage units or the addition of new ones without any break in service.

Storage Management Software

Any organisation that uses array storage from different vendors faces the problem that support staff have to use several storage management software interfaces to manage and monitor the disk arrays. This erodes productivity.

One solution is to use Microsoft's Windows Storage Server 2003 utilities, but this operating system is not supported by all storage vendors. Another solution is to adopt Storage Management Software products, which would allow storage managers to handle heterogeneous storage hardware from a single "enterprise window". There is no clear consensus about the relative merits of these approaches.

Obsolescent technologies

Much of the literature about digital repositories reads as if they change only by adding new materials to ever-increasing stores, but this can be only part of the truth. Whatever their technical infrastructures, change is inescapable, as is decay and its mitigation.

Factors in this include:

- Storage hardware is being continually developed. Today's 'state of the art' may be obsolescent in 5 years time and impossible to maintain in 20 years time.
- The skills used to maintain storage hardware and any associated software can similarly become hard to find, and even harder to train new staff in.

---

[12] Adapted from *Storage Outlook: what's on the storage horizon?*, Henry Newman, October 31 2003. www.enterprisestorageforum.com/technology/features/article.php/11192_3102791_2

[13] See for example *Information lifecycle management: destined for a bumpy ride*, Graham Titterington (Ovum), 2004. http://www.ovum.com/go/content/c,41487

- Electronic media are not as permanent as is often thought. Manufacturers may claim satisfyingly long lifetimes for their media [14] but practical experience suggests that a realistic figure for the life of a magnetic tape may be 15 years, and for a CD 20 years, all depending on original quality, storage, handling, and usage. And even if the media lifetime is longer, the hardware to read it may not be available.
- For many media, a small imperfection that appears after some time may make the whole medium unusable. A small part of a magnetic tape that has suffered chemical degradation may well damage the tape read/write head, so the whole tape will need to be discarded before any more damage is done.
- Standards are, for the best of reasons, in a state of constant flux. Today's DVD have been designed to be backward-compatible with CD-ROMs, and so can play old CD-ROMs, CD-I and CD-R disks, and video CDs, as well as new DVD-ROMs. But this position was not reached without some years of uncertainty about standards, and some parallel market-place confusion. Longer ago, the outcome of the Betamax vs. VHS standards 'war' reminds us that it is not always the technically 'better' option that wins out commercially. [15]

Storage technologies thus have to be considered as chronically obsolescent. Data storage is not a 'write once and forget' activity. To deal with the practical consequences, repositories will increasingly have to consider issues such as

- copying newly-received materials from an uncontrolled range of media and formats on to one selected medium, as the starting point for manageable storage
- periodic copying of data (selected by age, use, or other criteria) on to new media
- mechanisms that guarantee that the copies are faithful ones, down to the last bit.

Repositories will also need to replace storage components with new components using potentially different technologies. Factors influencing the decision to replace any storage include

- end of its active life
- its age
- the end of any maker's warranty
- the inability to obtain assured maintenance
- Total Cost of Ownership issues: for example, the lower operational costs of new components may more than compensate for acquisition costs being incurred sooner.

Emerging technologies

Absolute certainty about the way storage technology and the market will develop is not possible, but there are some straws in the wind.

- Storage manufacturers indicate that they see Serial Attached SCSI devices as the most likely future option, taking over from Serial ATA devices in due course.
- An interesting development is the idea of the Massive Array of Inactive Disks (MAID). With this, a rack-mounted disk array has all the disks powered down, with only the disk controller alive. When the application asks for an object, the controller powers up the appropriate disk drive(s), transfers the data, then powers the drive(s) down again. The potential operational economies of this development make it of great interest in the archive environment. [16]
- Tape drives cannot be ignored. Manufacturers claim that by 2010 they will have a tape that can store 1 Terabyte (TB) [17]: the standards being developed certainly point in that direction.[18]

---

[14] 1995 Kodak research on their writeable CDs, reported at http://www.cd-info.com/CDIC/Technology/CD-R/Media/Kodak.html, quoted a lifetime of 217 years under specified conditions.

[15] Some see this more as a battle of standards: see for example *Winners, Losers & Microsoft: Competition and Antitrust in High Technology*, Stephen E. Margolis and Stan Liebowitz, The Independent Institute, Oakland CA, 1999. ISBN 0 945999 80 1

[16] See, for example, *A MAID for your archives*, Mario Apicella April 30, 2004. http://www.infoworld.com/article/04/04/30/18storinsider_1.html

[17] One terabyte (TB) is confusingly defined, depending on context, as 1000 or 1024 gigabytes. Similarly, one gigabyte (GB) can be 1000 or 1024 megabytes.

[18] The LTO4 standard, expected at the end of 2004, promises 400GB storage, or up to 800GB after compression.

Case study: finding the best solution for the British Library

The British Library had to choose the best physical storage solution as part of developing its system to handle digital objects. Though driven by its own needs, perceptions, and present facilities, the problems that it had to solve are not peculiar to this one institution: many will apply to any large digital archive, especially an academic one.

Behind the problems are some general principles

- Some digital objects to be stored arise from purchased acquisitions and donations, and from digitising items already in the Library's collections. Other objects will be acquired as a result of Legal Deposit legislation [19] (the national library acting as the guardian of the national intellectual output). It will be managerially simpler not to distinguish materials by their source (which would imply having possibly many storage systems, with considerable cost consequences). It follows that the storage and handling methods adopted must be applicable to all materials.
- Speed of file handling is far less important than guaranteeing that a file will be neither lost nor damaged. To put it another way, preserving an object is paramount, providing access of some kind to it is secondary, and providing quick access (i.e. a response time of a few seconds) could be considered a luxury.
- The majority of the material stored may rarely or even never be used – but it is not possible to predict which material this will be. Although this is an assumption, it is based on many years experience with printed material received under Legal Deposit.
- Access to the material stored will be random.

The over-arching design principle

An initial understanding of IT storage technologies and the market itself led to some basic conclusions.

a) We could not rely on one storage technology, but would have to migrate through an unknown number of different technologies over time
b) We would have to be flexible in dealing with ever-changing vendors and types of vendor.
c) We foresaw an unknown (but large) storage requirement.

These in turn led to the principle that underlies our design and development work:

- **it is vital for the physical storage and retrieval of our digital objects to be independent of the technical properties and characteristics of the physical storage itself**.

Such a design should be able to accommodate many generations of physical storage implementations.

This led (at the end of 2003) to two immediate questions:

- what storage technology should we use initially?
- how should this design principle be put into practice in building our system?

We decided that the significance of what we were doing merited an extensive evaluation of UK suppliers. This was supported by research using numerous trade sources, and by the subscription-based services of Forrester Research.

Some prior assumptions

Storage growth

We assumed, for planning purposes, that about 300TB of storage would be required over the next 5 years. This was derived from a simple assumption of an initial 4TB, trebling each year (4TB in year 1, 12TB in year 2, 36TB in year 3, 108TB in year 3, and 324TB in year 5). This assumption is a 'best guess': such data as we have is not yet robust enough to give a better planning figure.

- Some statistics from the 'Voluntary Deposit of Electronic Publications' scheme are shown in Appendix A.
- Some of the storage will be used for web archiving. Both the number of web sites we will have to cover and the average size that may be expected are highly uncertain. The present estimate of the number of sites is 5 to 6 million, but this is an unrefined approximation.

---

[19] Although updated legislation became law in 2003, the enabling legislation to put it into practice has not yet been finalised, and it is unlikely that the law will come into effect until 2006.

The only firmer evidence of likely demand may come as the enabling legislation needed to put the new Legal Deposit law into practice is developed.

Different types of storage

Consideration of the system we would need led to our separating out 3 distinct storage needs.

- Preservation storage. Long-term permanent storage for the digital objects.
- Workspace storage. Temporary storage for objects while they are being ingested.
- Access storage. Temporary storage for objects that have been requested from preservation storage.

The discussion in this paper concerns only preservation and workspace storage.

Potential storage models

At the outset we identified 3 possible ways of building the storage we would need.

- Large monolithic storage. This could be scaled up to meet our needs, but offered features such as high availability and high access speed that we did not necessarily need, at a price we might not be able to afford.
- Commodity storage. Neither as fast nor as scalable as monolithic storage, but more readily available and liable to be less expensive.
- Assembly by ourselves of various components into the system we would need.

Interviews with suppliers

The companies seen between November 2003 and February 2004 are listed in Appendix B. We must stress that no company was excluded at this stage on technical or commercial grounds: any exclusions were due to our being unaware of the relevance of what the companies themselves could offer.

Our approach to suppliers was deliberately broad. The outline requirement put to them is shown in Appendix C.

Design principles

The discussions with suppliers took place over 3 months, during which time our thinking about the whole storage system was developing. Consideration of a number of issues, concerning

- the objects to be stored
- the need to manage costs
- the need to design for tolerance to disasters
- the mechanism of procurement
- the need to deploy and manage storage.

led to a number of system design principles, which any storage hardware would have to meet where appropriate.

We also agreed on the functions of the 'storage service layer', and developed our thinking about resilience and multi-site storage. This thinking in turn influenced our interpretation of what the suppliers were telling us.

- The storage system must be based on a multi-site (at minimum two-site) design that is not subject to obvious or plausibly likely common mode disaster scenarios.
- These nodes would be on geographically separate sites, for security reasons. [20]
- At least one alternative site must be capable of providing service (in terms of functionality if not in terms of performance) in the event of a disaster at another site and while restoration is in hand.
- We did not wish to be locked in to any one particular vendor, no matter how attractive their initial offerings might be. We needed to be free to react to market, technical and commercial changes, and to avoid complex procurement operations, none of which would be possible with a single vendor
- We needed to design the system so that it can accommodate storage sub-systems supplied by different vendors – i.e. we need to be able [21] to support heterogeneous, rather than homogenous, storage subsystems.

---

[20] The locations need careful thought. The case of the business that realised that its primary data centre and its data recovery location, though separate, were both in the Thames flood plain, may be apocryphal. It is still instructive.

However, the most far-reaching conclusion that we reached from these various perspectives was that

- This need for heterogeneous storage results in a design principle whereby other components in the storage service system work with a simple abstract (or neutral) model of storage sub-systems that can be readily applied to products from different vendors. Having this high level distinction between the storage service layer and the physical storage layer would allow us to procure physical storage independently of the rest of the system.

Functions of the storage service layer

The storage service layer would

- allocate unique persistent (vendor-independent) identifiers to objects
- bind each of these identifiers permanently to its object
- guarantee the authenticity and integrity of objects
- handle the reliable transport of each object as needed
- recover from any internal failure
- provide full service to users of another site if that other site was unavailable
- provide the means to integrate the physical storage with external systems.

Multi-site storage

Further consideration of the design of multi-site solutions looked at two options:

- master and standby systems, where only one system delivers live service at a time
- distributed peer systems, where both deliver service concurrently.

A principal limitation of master/standby approach is that only one half of the complete system is effectively delivering service at one time, and it is thus inherently inefficient: two sets of equipment are bought but only one is used to deliver service.

Distributed peer systems have the ability to deliver service concurrently from all equipment, and are thus intrinsically more efficient. However, one needs to ensure that the consistency of data is adequately managed especially for any read/write operations on the same data fields. (For example, without centralised locking it is inherently difficult to detect there are two concurrent and remote attempts to modify the same logical database record.) However, there is only one aspect of the behaviour of objects in the system that we were designing where this is a potential concern. This is the allocation of unique IDs to the objects within the system, which is a relatively simple matter to address.

The system we were designing looked well aligned with the design principles behind distributed peer systems. A simplified block diagram of the functional components (showing 2 sites, though any number within reason could be possible) is shown in Figure 2.

[21] The system has to be able to support heterogeneity. This does not preclude choosing to place future contracts with the same vendor.

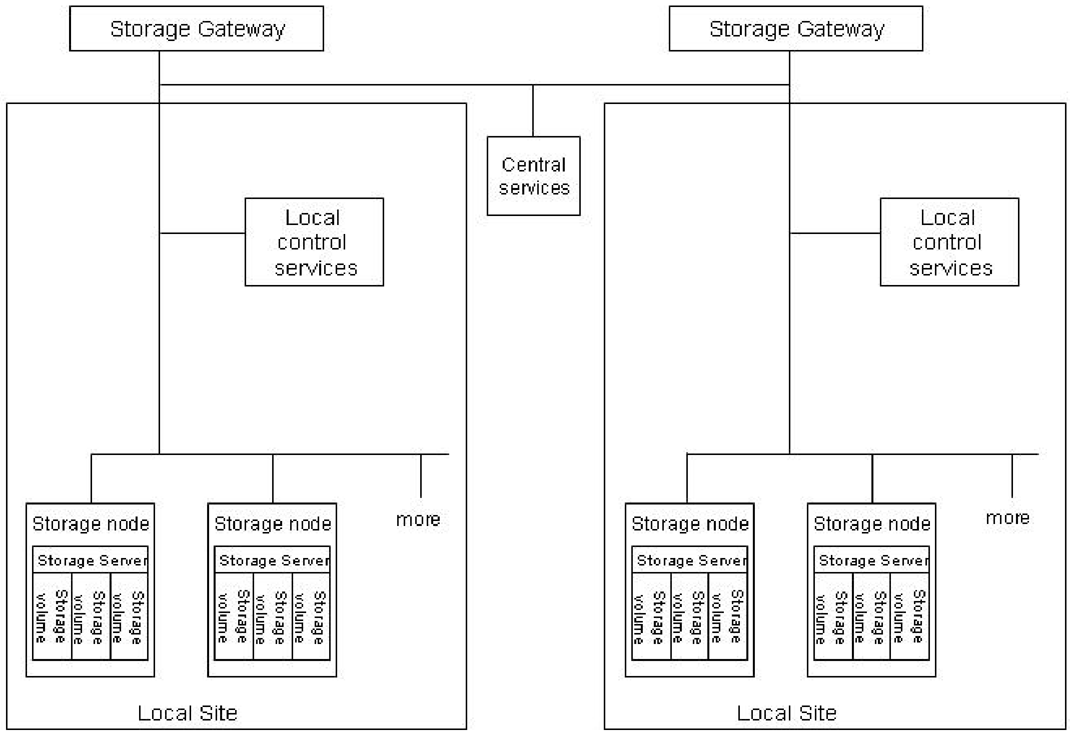

Figure 2. Two-site distributed storage system

Having made this choice, further design principles become apparent.

- The functional behaviour of each site should be identical in terms of logical architecture, interfaces, and capability (except when being upgraded as discussed later).
- However, there should be no need for the detailed physical equipment to be identical on each site. For example, a replacement volume installed after hardware failure at one site may have a different capacity from the original.
- Each site can be autonomous, in that it can make local decisions about where objects are stored locally without reference to any other sites.
- It must be possible to deliver all services from one site when the other site(s) are unavailable (whether for planned or unplanned reasons).
- A total upgrade must be achieved, with service delivered throughout, by applying software upgrades or new releases to each site in turn.
- Interoperability between sites is important, so one site must be able to operate on a different software version from the other site(s). This principle may need to be applied only during the period of a software installation: during the installation itself, sites will be using different versions, but afterwards all sites will be on a new common version.
- A multi-site design, though prompted by disaster tolerance, also has a wider bearing on resilience. There are at least two live independent instances of each object.
- This solution is preferred to having a single store with high built-in redundancy, resilience, and recovery facilities.

Conclusions from the supplier presentations

The supplier briefings were intensive and allowed us to assemble and check a number of conclusions.

The storage industry

- The storage industry is going through a large amount of 'churn', with company acquisitions and mergers, and changes to business models.
- Many small start-up companies are introducing innovative new hardware solutions, and either surviving, being purchased, or disappearing.
- Replacement disk technologies require venture capital investment, which can be scarce.
- The well-established vendors of enterprise storage hardware are introducing commodity Serial ATA (SATA) disks into their arrays, reducing costs in order to be able to sell products in the Small to Medium Enterprise (SME) arena.
- The same technology allows smaller storage hardware companies to increase the capacity of commodity-priced Network Attached Storage (NAS) devices, broadening their usage from that of a basic storage solution and pushing sales up into the SME arena.
- Revenue opportunities for large storage companies have moved away from the sales of storage hardware and more into the realms of associated services and software licensing, support and maintenance.
- Revenue opportunities for smaller or 'start-up' storage hardware companies lie in providing ever-increasing capacity for reasonable costs.
- There are only a few manufacturers of the actual disk drive units: the different vendors are differentiated by the way in which they integrate the units into large-scale storage systems, and the services they supply with these systems.
- Microsoft is strongly advocating the use of Windows Storage Server 2003 as a storage operating system.
- The dominant thrust in the market is to deliver performance within a highly resilient single cluster, supplied by a single vendor. This means that a working relationship has to be set up with this one vendor for the medium-term (several years). This position is at odds with our need to think very long-term (centuries), while our need to minimise Total Cost of Ownership also dictates against forming too close a relationship with any particular vendor, no matter how attractive this may seem in the short term.

Storage technology

- The monolithic storage companies have established the latest 'Holy Grail' in the storage industry, that of 'Information Lifecycle Management'. Here, companies offer suites of storage hardware and software which allow the purchaser to manage their data 'from the cradle to the grave', controlling the data migration from fast, high-performance storage to lower-performance but more cost-effective storage as the value of the information changes and as use decreases. This view of storage is at odds with our own view.
    - The lifetime of digital objects is not limited. Deletion is not an option.
    - It is not our role as a library to judge the intellectual quality of all the objects and build content management rules to exploit such judgements (though there is an undoubted role for Records Managers and the like to help define these policies within their own organisations). The most we would do would be to build links between related objects.

  To this extent there is a mismatch between our requirements of permanent secure storage and what much of the industry is concentrating on developing.
- Of the three storage models considered, that of building our own storage system from bought-in components was judged to have a high risk of failure (though not necessarily immediately), and was quickly ruled out. The two storage models left in contention were worth further consideration.
- No matter what technology is adopted, the equipment will have a limited lifespan, governed by a combination of the maker's warranty period and our internal policies on hardware write-off. There will thus be an ongoing programme for the replacement of equipment that has reached the end of its working life. This too leads to a reluctance to become too heavily dependent on or involved with a single vendor.

- Serial ATA (SATA) disk technology seems well-suited for the preservation storage of digital items. It is a new technology building on an established technological foundation [22], offering an appropriate combination of cost, volume, access speed, and manageability. It is a technology that seems to have a developing market which some vendors are eager to exploit (i.e. we do not see it as a technology for a niche market). We had some concerns about the maturity of the SATA standard, in particular where it relates to data resilience features of its disk controllers. There seemed to be some doubt about how manufacturers were implementing solutions which met the standard as it was at the time, but we are aware of intensive work on the standard that is being carried out.[23]
- Conversely, the extremely fast, highly resilient cached disk arrays available from some vendors are not an ideal fit with our preservation storage requirements. A significant number of objects held in the storage system will rarely be accessed, so the ability to deliver them faster offers little value. As a general guide, products that are capable of delivering twice the performance may increase costs up five-fold. We need storage that must be scaleable in capacity and yet still be affordable: maximising performance with a disproportionate increase in costs is inappropriate. However, we do recognise that there will be practical minimum performance levels that must be met.
- Another reason that the enterprise-level solutions offered do not meet our needs is that the additional resilience, such as multiple data paths, controllers, etc, does not offer us as much value as it would in a conventional enterprise solution. These features are typically deployed where there is no second on-line system. Since there are at least three[24] independent instances of each object, and an object will be delivered from a remote site if a local site is unable to deliver it, expensive additional resilience adds little additional value. However, RAID technology (specifically RAID 5 and derivatives which offer error detection and on-line correction) is in effect a low cost commodity and seems very appropriate.
- The high-speed bandwidth and transport capabilities afforded by Fibre Channel Storage Area Networks similarly seem to be a poor fit for our preservation storage requirements, since speed is not of the essence. They were however a good fit with our needs for workspace storage, where potentially large objects would be ingested for initial processing.
- The maturity of the standards behind the iSCSI protocol suggest that it may be a better storage networking solution for us in the medium term, as overall storage capacity increases.

Object retrieval

- The majority of digital items ingested into our system may be retrieved rarely or even never, but they must be preserved in perpetuity. Once items number in the millions, the retrieval of say 15-20% of them over a number of years will constitute a significant retrieval requirement.
- Given the likely retrieval response times for stored items required by readers in our Reading Rooms, tape is not an ideal storage medium. We would not envisage offering a service with sub-second response times, but nor can we see a time when tape will be a viable way of offering response times at the under-10-second level.

Procurement

- It is perfectly feasible to procure the storage we need for digital objects on a rolling basis, in order to take advantage of storage cost reductions and increases in individual disk media capacities. In practice we are likely to conduct a procurement exercise on an annual, or more frequent, basis with the option to call off additional purchases as required during the period of any contract.
- Although this rolling procurement will result in an heterogeneous storage layer, we have designed a storage application layer which is abstracted from the physical storage layer. The storage service layer should contain no components that depend on properties of the actual storage used.

---

[22] The standardized SATA specification replaces a 15 year old technology with a high-speed data transfer technology that has an expected future of 10 years.

[23] The SATA International Organization provides the industry with guidance and support for implementing the SATA specification. It was officially formed in July 2004 by incorporating the previous Serial ATA Working Group. See http://www.sata-io.org/

[24] As initially implemented, these are two sites plus the ‘Dark Archive’, discussed below.

The 'Dark Archive'

Although the discussion above has centred on a general multi-site storage model (one that could be described as 'using multiple autonomous independent peer clusters that cross-synchronise'), in practice this is to be implemented as a two-site model. Questions of resilience still remain: what happens if both sites fail, or are found to be prone to a common mode of failure? What is the way back from total loss?

Whilst it is possible to obtain commercial disaster recovery solutions for common equipment configurations, we believed that it was not feasible to obtain such solutions for systems storing many hundreds of terabytes. We concluded that we had to incorporate the need for disaster recovery into the design of our system. Since the weak point is that the sites will be identical, or nearly so, we concluded that we would need an independent instance of the data in a totally separate repository using totally different technology. This concept has become known as the 'dark archive'.

The full details of this archive have yet to be worked out, but possibilities include the use of write-once media, and avoiding using any of the existing software components, other than to receive data. The archive is a strong candidate for being hosted by a third party. If nothing else, it would need to be geographically separate from any of the storage sites.

In a two-site system, this would give three independent instances of each object. We considered that this was sufficient to fulfil the need for resilience. It follows that there is no additional need for objects to be subject to conventional backup processes, and the considerable costs that they entail.

Final procurement

Based on the considerations outlined, the British Library moved to procure storage in June / July 2004. Suppliers were asked to propose a solution covering

- 2 discrete disk arrays each of 5 terabytes, for Preservation Storage
- 2 discrete disk arrays each of 1 terabyte, for Workspace Storage
- details of solutions for connectivity with existing servers
- any other additional software.

The full Technical Requirements are given in Appendix D.

We told prospective suppliers what our mandatory and desirable requirements were, and what information we expected them to supply. In this way we hoped that there would be no doubt about what we wanted or how we would evaluate any proposals.

Choice of vendor

Our procurement was conducted using the procedures appropriate for a UK Government body. After careful consideration of the sixteen responses that we received we awarded the tender to VSPL, who proposed a solution using JetStor disk arrays.[25] The cost quoted was well under £100,000. To put it another way, the acquisition cost (allowing for the fact that the digital objects being stored are all duplicated) was about £9 per effective GB.

This solution has a good reputation amongst storage managers, as having good performance and reliability. We believe that it will meet our Total Cost of Ownership criteria. We recognise however that the storage management software is specific to the JetStor range and does not comply with some storage industry standards.

The procurement was instructive in that the gap between the acquisition costs of this solution and those of others offered was clear enough for any variances in our approximate costings of other factors to be of little relevance. A subsequent purchase of other storage (in this case Hewlett Packard) showed us just how volatile the purchase costs of storage can be.

Practical experience

It is too early to draw any long-term conclusions about the wisdom of our design choices or our first choice of storage vendor. We also have no experience as yet in managing storage from multiple vendors. However, we can at least say the following.

- The vendor's promise of very quick delivery was, with one non-critical exception, fully honoured.

[25] See http://www.vspl.co.uk/jetstor.htm

- Setup and configuration presented no great difficulties to staff who were well versed in using other storage hardware.
- The units have been successfully load-tested, but they have not at the time of writing been used in anger.

Future procurements

- We were aware that one component of the solution was a NAS component attached to each storage controller. There is a scalability problem here which could become problematic: further storage purchases will bring in more of these components, which will have to be maintained separately rather than as a single group.
- The storage management software was known in advance to be limited. A separate procurement has been started to purchase such software to manage not only the JetStor units but potentially units from other vendors. It should also deal with the NAS problem outlined above.

Application of the case study to other institutions

This paper has described storage technology issues and how the British Library responded to its perception of its needs and the state of the storage market. The experience is relevant to other institutions for the reasons and with the limitations that follow.

The problems are not unique

Libraries and archives have a traditional role of preserving and allowing access to materials over (in cultural terms) long times-scales, i.e. for many decades, even centuries. The digital age is not going to alter this role substantially, but it is making it more difficult to carry out because of

- the explosion in the amount of material, and
- the in-built obsolescence of the storage media and the technologies to exploit them.

Our requirements differ from those of other institutions because of our position vis-à-vis Legal Deposit in the UK[26], which imposes certain legal obligations on us. However, similar obligations are accepted by other repositories, either as outcomes of other external constraints or as matters of internal policy. Our long-term needs are far from unique.

The volumes of material that we are planning for may appear large, but we will be far from the only digital repository that has to manage the volumes envisaged.

The solutions we have adopted do not necessarily give quick access to any stored object. We have made it clear that the security and long-term preservation of any object is of far greater importance than speed of access to it.

A repository may have existing institutional policies that exclude the solutions described here at present, but these policies will have to be adapted as technologies develop, and these solutions may fall within future constraints.

The technology options are the same for all

The technology available is changing constantly. Data storage vendors are very active with new standards and solutions, and there is no prospect in sight that any repository can put its own storage technology issues to one side for any length of time. The culture is one of 'constant revolution', and the solutions adopted today have to allow for that.

Storage vendors will continue to have a range of products, not all of which are suited to non-commercial activities such as digital repositories. Our experience has led to what should be a less expensive solution than a more mainstream commercial one.

For a depository with different storage volumes, the storage solutions adopted may still be the same in essence even if they differ in actual numbers. The solutions are designed to be flexible, and we are confident that they can handle data volumes an order of magnitude less or greater than what we have planned for.

---

[26] We are actively involved with the other Legal Deposit libraries in finding answers to the problems that the extension of Legal Deposit to electronic materials brings.

Conclusion

The British Library is building a large-scale digital preservation system (the Digital Object Management System). Some key features of this are

- it will have a large capacity, which at present is difficult to predict
- it must ensure that the objects are stored securely
- the life-time of objects must be taken as 'forever'
- access to stored objects will be random
- accesses for most objects may well be low
- speed of access to objects does not need to be sub-second, but should be a few seconds
- it must be resilient when faced with planned or unplanned disruption
- life-time costs must be as low as possible, given all the other features.

This paper has described an extensive review of the storage market. This offers basically two different solutions, neither of which meets our needs.

- High-end enterprise storage. This is expensive, and offers many unnecessary features. It also involves lock-in to a vendor, which may not be a long-term advantage.
- Lower-end commodity storage. This is less expensive, but it does not give enough resilience for our archival requirements.

We have evolved a system design where the physical storage and retrieval of our digital objects is independent of the technical properties and characteristics of the physical storage itself. This allows us to assemble a distributed storage system with commodity storage units from (potentially) a number of vendors. The system design also ensures resilience.

The general requirements that we have, and the state of storage technology and the storage market-place, mean that the conclusions we have drawn are also relevant to any digital archive or repository with a long-term (multi-decade) view and large (multi-terabyte) anticipated storage volumes.

Further contact

For further information about the British Library Digital Object Management Programme, please contact:

Richard Masters,
DOM Programme Manager,
eIS,
British Library,
Boston Spa,
Wetherby,
West Yorkshire,
LS23 7BQ,
UK

Phone 01937 546888
Fax 01937 546872
Email richard.masters@bl.uk

Glossary

ATA	***A**dvanced **T**echnology **A**ttachment.* A disk drive implementation where the control of data transfer between the computer and the disk drive is part of the disk drive's functionality.
Fibre Channel	A serial data transfer architecture standardized by ANSI. The most prominent Fibre Channel standard is *Fibre Channel Arbitrated Loop (FC-AL)*. FC-AL was designed for new mass storage devices and other peripheral devices that require very high bandwidth. Using optical fibre to connect devices, FC-AL supports full-duplex data transfer rates of 100MBps. FC-AL is compatible with, and is expected to eventually replace, SCSI for high-performance storage systems.
IP	***I**nternet **P**rotocol.* IP specifies the format of data packets and the addressing scheme for them. Most networks combine IP with a higher-level protocol called *Transmission Control Protocol (TCP)*, which establishes a virtual connection between a destination and a source. IP by itself is something like the postal system. It allows you to address a package and drop it in the system, but there's no direct link between you and the recipient. TCP/IP, on the other hand, establishes a connection between two hosts so that they can send messages back and forth for a period of time.
iSCSI	***I**nternet **SCSI***, an IP-based standard for linking data storage devices over a network and transferring data by carrying SCSI commands over IP networks.
LTO	***L**inear **T**ape **O**pen.* This technology combines the advantages of linear multi-channel bi-directional tape formats with enhancements in the areas of timing, data compression, track layouts, and error correction, so as to maximize capacity and performance.
Mass storage	Refers to various techniques and devices for storing large amounts of data. Mass storage devices include all types of disk drives (magnetic disks and optical disks) as well as tape drives. They retain data even when the power supply is turned off. Mass storage is normally measured in kilobytes (1,024 bytes), megabytes (1,024 kilobytes), gigabytes (1,024 megabytes) and terabytes (1,024 gigabytes). Sometimes the multiplier of 1,000 is used instead of 1,024.
RAID	***R**edundant **A**rray of **I**nexpensive **D**isks*. RAID technology was initially devised as a way to distribute large amounts of data across an array of small disks. The purpose of RAID is twofold • to avoid reliance on a single disk, • to implement relatively low-cost disks. In a RAID application, even if a disk drive crashes, the host will remain up and running. With today's "hot-swappable" disk drives, the failed drive can be replaced while the server remains fully operational. There are several established ways of implementing RAID, using different minimum numbers of disks and different combinations of disk mirroring, copying data over several disks, and error checking. RAID level 1 and RAID level 5 are amongst the commonest implementations.
SCSI	***S**mall **C**omputer **S**ystem **I**nterface.* A parallel interface standard used by PCs, Apple Macintosh computers, and many UNIX systems for attaching peripheral devices to computers. SCSI interfaces provide for faster data transmission rates (up to 80 megabytes per second) than standard serial and parallel ports. In addition, many devices can be attached to a single SCSI port, which makes SCSI more than simply an interface. Although SCSI is an ANSI standard, there are at least 9 variations of it, so two SCSI interfaces may be incompatible. For example, there are several types of SCSI connectors.
Serial across SCSI, also Serial Attached SCSI (SAS)	SAS is a performance improvement over traditional SCSI as it enables multiple devices (up to 128) of different sizes and types to be connected simultaneously with thinner and longer cables; it supports a data transfer rate of 3.0 gigabits per second.
Serial ATA (SATA)	Often abbreviated *SATA* or *S-ATA*, Serial ATA is an evolution of the Parallel ATA physical storage interface. It is a serial link -- a single cable with a minimum of four wires creates a point-to-point connection between devices. Transfer rates for Serial ATA begin at 150MBps. One of the main design advantages of Serial ATA is that the thinner serial cables facilitate more efficient airflow inside a physical device and also allow for smaller chassis designs.
Storage controllers	The storage control unit that performs command decoding and execution, host data transfer, error detection and correction, and overall device management. It may also control the operation of any RAID technology and performance optimisation for multiple storage devices.
Storage Management Software	Software to manage data storage. Performs tasks such as data backup and recovery, storage resource management (including moving infrequently moved objects to cheaper storage), storage network management, real-time monitoring, and storage management control.
Storage Resource Management	This has evolved from being a simple data gathering and reporting tool to being a proactive application to control the allocation of data storage, storage structures, and data integrity.
TCP	***T**ransmission **C**ontrol **P**rotocol*. TCP enables two hosts to establish a connection and exchange streams of data. TCP guarantees delivery of data and also guarantees that packets will be delivered in the same order in which they were sent.

Appendices

Appendix A. VDEP Statistics

The Voluntary Deposit of Electronic Publications scheme has been running since 2000 [27]. This scheme encourages UK publishers to deposit with the British Library electronic publications issued on discrete physical digital media such as magnetic tapes, magnetic disks or, more commonly, optical disks of some kind, such as CD-ROM or DVD. Other materials are also deposited: for example, many files have been received as e-mail attachments. Participation in the scheme is totally voluntary.

By the end of 2004

- There were approx, 180,000 objects. Some 1300 serial titles formed multiple objects within this total.
- The storage volume used was 1.45 TB
- The average weekly growth in 2004 was 12.5 GB per week
- Object sizes ranged from 1 KB to over 640 MB, averaging approx. 8.4 MB each
- The breakdown of file-types was:

.txt	43%
.htm, .html	34%
.pdf	9%
.jpg	3%
.ZIP	3%
.xls	2%
.doc	2%
37 others	4%

Appendix B. Suppliers interviewed

This list shows the suppliers seen, with company URLs. (All links were active at 31st January 2005.)

Supplier	Web address
3Pardata	www.3pardata.com/
ADIC	www.adic.com/adicHomePage.jsp
AppIQ	www.appiq.com/
BlueArc UK Ltd	www.bluearc.com/
CA	www3.ca.com/Solutions/Solution.asp?id=370
Chaparral Network Storage	www.chaparralnet.com/
CommVault Systems	www.commvault.com/
Creekpath	www.creekpath.com/
Datacore	www.datacore.co.uk/index.asp
Dell (with EMC)	www.dell.com/ www.emc.com/
Fujitsu Softek	www.fel.fujitsu.com/home/v3__default.asp
Hawke Systems	www.hawke.co.uk
Hitachi Data Systems (with Morse)	www.hds.com/
HP	welcome.hp.com/country/uk/en/prodserv/storage.html
hps	www.hps.co.uk/
hps (with Cloverleaf Comms)	www.hps.co.uk/ www.cloverleafcomm.com/
IBM (with Sagitta Performance Systems)	www.ibm.com/uk/ www.sagitta-ps.com/
InTechnolology	www.intechnology.co.uk/html/reseller/ips/intro.asp
Iomega	www.iomega.com/global/index.jsp
McDATA UK	www.mcdata.com/
Network Appliance	www.netapp.com/
Plumdata	www.plumdata.co.uk
Q Associates	www.qassociates.co.uk/index.htm
SGI	www.sgi.com/
Source Consulting	www.sourceconsulting.com

[27] For details and further links see http://www.bl.uk/about/policies/legaldeposit.html#elec

Storagetek	www.storagetek.com/
SUN	www.sun.com/
Tek-Tools Inc	www.tek-tools.com/
Veritas	www.veritas.qassociates.co.uk/index.htm
VSPL	www.vspl.co.uk

Appendix C. The formal requirement put to the suppliers.

The British Library intends to procure storage hardware, management and other related software in the second quarter of 2004. Detailed requirements still need to be developed, however currently we assume that equipment purchased during this phase will be used as a pilot system to prove the effectiveness of a scalable storage solution for preserving and providing access to digital objects.

It is difficult to create estimates of initial and ongoing storage requirements, and therefore at this stage the team are proposing a procurement process which will result in a storage system with sufficient flexibility and scalability to address this current uncertainty. At this time we envisage a comparatively modest initial storage purchase, with regular future purchases of additional storage to cater for rapid growth commencing in early 2005.

We anticipate that the storage content will consist of a significant number of objects which will be rarely accessed but which are stored on-line, and a small number of objects which are frequently accessed. We assume that some form of cached and responsive storage will be needed for the latter, and vendors are invited to cover any relevant products as part of their briefing presentation.

We anticipate our storage needs will be met by one (or a combination) of the following scenarios:

- Purchase of a complete solution from a single vendor which meets storage hardware, storage management and digital object integrity application requirements; alternatively the provision of a hardware and software solution comprising the same functionality by an Independent Solution Vendor
- Purchase of storage hardware directly from a major storage vendor; purchase of storage management software from the same vendor, if capable of managing heterogeneous storage components, or from an alternative vendor if not. Additional storage may subsequently be purchased from vendors providing lowest Total Cost of Ownership when measured throughout the storage lifecycle. (Note that from a strategic perspective this assumes a heterogeneous hardware storage layer developing over time.)
- The in-house construction of a storage system using commodity storage components procured separately from a range of vendors, and the separate procurement or internal development of storage control and management software.

An important selection criterion for the Library will be the Total Cost of Ownership of storage components, applied to the full component lifecycle.

Appendix D. Technical requirements for suppliers

Suppliers were asked to propose a solution covering

- 2 discrete disk arrays for Preservation Storage, each with a capacity of 5 terabytes (5TB) when formatted as RAID-5. We explained that we assumed that these would comprise Serial or Parallel ATA disks attached to one or more SATA (or other) controllers within a rack mount enclosure.
- 2 discrete disk arrays for Workspace Storage, each with a capacity of 1 terabyte (1TB) when formatted as RAID-1. These are for temporary workspace storage of digital items. We explained that we assumed that these would be high performance disk arrays, using Fibre Channel (FC) or UltraSCSI disks, attached to one or more Fiber Channel or Ultra-SCSI controllers within a rack mount enclosure.

To support this, suppliers were required to

- specify a storage connectivity solution, which would allow the existing system servers to access array storage.
- provide details of the capabilities of any array management software which was bundled with the specified array solutions. We recognized that vendors might wish to specify additional software components as part of their solution providing, for example, a consolidated view of storage from the application level, or significantly reduced storage administration resource effort.
- identify any training which was specific to the solution components and which was essential for the effective use and management of the overall solution.
- provide itemized costs for work required to install, configure and commission solution components.
- provide costs and details of all relevant warranty and maintenance offerings.
- provide details of reference sites.